Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster)
Volume 04, No. 3 (2015), hal 305 – 312.

PERBANDINGAN IMPUTASI *MISSING DATA* MENGGUNAKAN METODE *MEAN* DAN METODE ALGORITMA *K-MEANS*

Mukarromah, Shantika Martha , Ilhamsyah

INTISARI

Missing data adalah informasi yang hilang atau tidak tersedia untuk sebuah obyek. Missing data merupakan masalah yang sering dijumpai dalam penelitian, keberadaan missing data dapat mengganggu analisis yang akan dilakukan. Salah satu cara yang dapat dilakukan untuk menangani missing data adalah dengan mengisi missing data dengan nilai-nilai yang mungkin berdasarkan informasi yang tersedia pada data atau dikenal dengan imputasi. Mean dan Algoritma K-Means merupakan metode yang dapat digunakan untuk imputasi missing data. Imputasi dengan metode Mean mengisi missing data dengan rata-rata nilai yang diketahui pada suatu variabel, sedangkan imputasi dengan metode Algoritma K-Means mengisi missing data dengan centroid yang sesuai dengan letak missing data berada. Dalam penelitian ini, dibandingkan kedua metode imputasi tersebut, yaitu dengan membandingkan nilai MSE (Mean Square Error) yang diperoleh masing-masing metode. Semakin kecil nilai MSE maka semakin kecil kesalahan hasil imputasi. Dari pengujian imputasi yang telah dilakukan yaitu pada data yang mengandung 10%, 20% dan 30% missing data, didapat bahwa secara rata-rata imputasi missing data menggunakan metode Algoritma K-Means dengan 2 cluster, 3 cluster dan 4 cluster menunjukkan hasil yang lebih baik dibanding metode Mean.

Kata Kunci : *missing data, imputasi, Algoritma K-Means*

PENDAHULUAN

Salah satu permasalahan yang sering terjadi pada penelitian adalah data hilang atau biasa dikenal dengan *missing data*. *Missing data* adalah informasi yang tidak tersedia untuk sebuah obyek [1]. *Missing data* dapat disebabkan karena kesalahan sistem seperti tidak adanya respon terhadap sensor atau perangkat penerima input, dapat pula disebabkan karena kesalahan manusia seperti kelalaian dalam pengumpulan data, ketidakmampuan responden dalam memberikan jawaban yang akurat atau responden tidak berkenan memberikan jawaban yang akurat [2]. *Missing data* merupakan hal yang tidak diinginkan oleh peneliti, karena dengan adanya *missing data* maka data tidak dapat dianalisis dengan baik.

Ada beberapa cara yang dapat dilakukan untuk menangani *missing data* seperti : *Listwise deletion*, *Pairwise deletion* dan Imputasi [3]. *Listwise deletion* yaitu dengan cara menghapus kasus (obyek) yang mengandung *missing data*. *Pairwise deletion* yaitu dengan cara menghapus *missing data*, sehingga yang dianalisis hanya nilai-nilai yang tersedia saja. Imputasi yaitu dengan cara mengisi *missing data* dengan nilai yang mungkin berdasarkan informasi yang tersedia pada data [4]. Imputasi merupakan pilihan penanganan *missing data* yang lebih baik dibanding *listwise deletion* dan *pairwise deletion*, karena untuk mengumpulkan data membutuhkan waktu yang lama dan biaya yang besar.

Dalam penelitian ini dibandingkan dua metode imputasi yaitu imputasi dengan metode *Mean* dan imputasi dengan metode Algoritma *K-Means* untuk mengetahui metode yang lebih baik dalam menangani *missing data*. Dalam pengujian imputasi ini digunakan data lengkap, kemudian dari data lengkap tersebut dilakukan penghilangan nilai dengan mekanisme MCAR (*Missing Complete At Random*) dengan persentase 10%, 20% dan 30%. MCAR merupakan salah satu mekanisme *missing data*. *Missing data* dengan mekanisme MCAR berarti bahwa *missing data* terjadi secara acak, yaitu jika kemungkinan nilai yang hilang pada variabel Y tidak berhubungan dengan nilai-nilai di variabel lain dan tidak berhubungan dengan nilai–nilai di variabel Y itu sendiri [5]. Setelah dilakukan

penghilangan nilai kemudian dilakukan imputasi terhadap nilai-nilai yang hilang menggunakan metode *Mean* dan metode Algoritma *K-Means*. Jumlah *cluster* yang digunakan untuk imputasi menggunakan metode Algoritma *K-Means* adalah 2, 3 dan 4 *cluster*. Setelah dilakukan imputasi dari kedua metode dilakukan evaluasi hasil imputasi yaitu dengan menghitung *Means Square Error* (MSE) [2]. MSE disini digunakan sebagai ukuran untuk mengetahui ketepatan hasil imputasi. Semakin kecil nilai MSE maka semakin kecil kesalahan hasil imputasi. Jika h merupakan banyak imputasi, r_s merupakan data asli (data yang dihilangkan yang akan diimputasi) dan o_s merupakan data hasil imputasi, maka nilai MSE dari hasil imputasi dapat ditentukan dengan persamaan sebagai berikut:

$$MSE = \frac{\sum_{s=1}^{h}(r_s - o_s)^2}{h} \quad (1)$$

Dalam pengujian imputasi ini dilakukan 5 kali replikasi untuk tiap-tiap persentase *missing data* sehingga dari 5 kali replikasi akan didapatkan nilai rata-rata MSE, kemudian dibandingkan nilai rata-rata MSE dari kedua metode.

IMPUTASI DENGAN METODE *MEAN*

Mean merupakan salah satu metode imputasi yang paling umum digunakan. Imputasi dengan metode *Mean* mengisi *missing data* dalam suatu variabel dengan rata-rata dari semua nilai yang diketahui pada suatu variabel [6]. Imputasi dengan metode *Mean* memiliki kelemahan yaitu mengurangi varians pada variabel [7], karena nilai yang diisikan adalah sama untuk setiap variabel. Berikut adalah contoh pengerjaan imputasi *missing data* dengan metode *Mean.*

Tabel 1. Data untuk contoh imputasi dengan metode *Mean*

Obyek	**Data yang mengandung *missing data***		**Data yang sudah di imputasi**	
	v1	**v2**	**v1**	**v2**
A	5,1		5,1	3,23
B	4,9	3	4,9	3
C		3,2	4,88	3,2
D	4,6	3,1	4,6	3,1
E	5		5	3,23
F	5,4	3,9	5,4	3,9
G	4,6	3,4	4,6	3,4
H	5		5	3,23
I	4,4	2,9	4,4	2,9
J	4,9	3,1	4,9	3,1

Missing data di obyek C pada v1 diisi dengan rata-rata dari semua nilai yang diketahui di v1 yaitu $\frac{5{,}1+4{,}9+4{,}6+5+5{,}4+4{,}6+5+4{,}4+4{,}9}{9} = 4{,}88$. Selanjutnya untuk mengisi *missing data* di obyek A, E dan H di v2 adalah dengan rata-rata dari semua nilai yang diketahui di v2 yaitu $\frac{3+3{,}2+3{,}1+3{,}9+3{,}4+2{,}9+3{,}1}{7} = 3{,}23$. Dari Tabel 1 dapat dilihat bahwa nilai yang diisikan untuk setiap *missing data* disuatu variabel adalah sama yaitu di v1 dengan 4,88 dan di v2 dengan 3,23. Oleh karena itu semakin banyak persentase *missing data* pada suatu variabel maka akan semakin mengurangi varians dalam suatu variabel data.

IMPUTASI DENGAN METODE ALGORITMA *K-MEANS*

Algoritma *K-Means* merupakan metode pengelompokan data non hirarki, jumlah kelompok yang akan dibentuk sudah terlebih dahulu diketahui dan ditentukan jumlahnya. Algoritma *K-Means* berusaha mempartisi obyek kedalam satu atau lebih *cluster* atau kelompok berdasarkan karakteristiknya, sehingga obyek yang mempunyai karakteristik yang sama dikelompokkan dalam satu *cluster* yang sama dan obyek yang mempunyai karakteristik yang berbeda dikelompokkan kedalam *cluster* yang lain [8]. Selain digunakan untuk pengelompokkan data, algoritma *K-Means* juga dapat digunakan sebagai salah satu penanganan *missing data* dengan imputasi. Sebelum menggunakan Algoritma *K-Means* yang perlu diperhatikan adalah satuan data. Jika data mempunyai satuan yang berbeda secara signifikan maka harus dilakukan proses standarisasi dengan mengubah data yang ada ke Z-Score [1]. Berikut adalah rumus Z-Score :

$$Z_i = \frac{X_i - \bar{X}_N}{S} \tag{2}$$

dimana : Z_i = Nilai variabel hasil standarisasi untuk obyek ke-i
X_i = Nilai variabel untuk obyek ke-i
$\bar{X}_N$ = Rata-rata
s = Standar deviasi

Adapun langkah-langkah imputasi *missing data* menggunakan Algoritma *K-Means* adalah [9]:

1. Seluruh obyek (termasuk obyek yang mengandung *missing data*) dikelompokan dengan Algoritma *K-Means*

   Adapun langkah-langkah Algoritma *K-Means* adalah sebagai berikut :

   1. Tentukan jumlah *cluster* (k) yang dibentuk dan jumlah iterasi maksimum (maxter).
   2. Tentukan *centroid* awal (C_0) secara random dari obyek-obyek data komplit yang tersedia sebanyak k *cluster*.
   3. Hitung jarak setiap obyek ke masing-masing *centroid*. Perhitungan jarak yang digunakan adalah jarak euclid, jarak euclid dihitung menggunakan rumus sebagai berikut :

      $$d(x_i, v_p) = \sqrt{\sum_{j=1}^{m}(x_{ij} - v_{pj})^2} \qquad ; j = 1, 2, 3, \dots, m \tag{3}$$

      dimana : $d(x_i, v_p)$ = jarak antara obyek ke-i dan *centroid cluster* ke-p
      x_{ij} = data pada obyek ke-i pada variabel ke-j
      v_{pj} = *centroid cluster* ke-p pada variabel ke-j
      m = banyaknya variabel (kolom)
   4. Kelompokkan setiap obyek berdasarkan jarak terdekat antara setiap obyek dengan masing-masing *centroid*.
   5. Lakukan iterasi (t), tentukan posisi *centroid* pada iterasi ke-t (C_t) dengan rumus sebagai berikut:

      $$v_{pj} = \frac{1}{N_p}\sum_{i=1}^{N_p} x_{ij} \tag{4}$$

      dimana : v_{pj} = *centroid cluster* ke-p pada variabel ke-j
      N_p = banyak/jumlah data yang menjadi anggota *cluster* ke-p
      x_{ij} = data pada obyek ke-i pada variabel ke-j
   6. Ulangi langkah 3 jika posisi $C_t \neq C_{t-1}$ atau jika t < maxter. Jika posisi $C_t = C_{t-1}$ atau jika t = maxter, maka proses perhitungan dihentikan dan didapatlah kelompok data.

2. Isi *missing data* dengan *centroid yang* sesuai dengan letak *missing data* berada.

Berikut adalah contoh pengerjaan imputasi *missing data* dengan metode Algoritma *K-Means*

Tabel 2. Data untuk contoh imputasi dengan metode Algoritma *K-Means*

Obyek (konsumen)	Usia (tahun)	Income (Rp)	Koran (Jam)	Tv (Jam)
A	25		10	20
B	26	750000	11	18
C		300000	5	
D	40	750000	9	15
E	35	500000		11
F	30	450000	6	14
G	25	250000	6	
H		400000		20
I	26	450000	5	
J	21	300000	3	15

Sebelum mengelompokan data dengan Algoritma *K-Means*, data terlebih dahulu distandarisasi dengan rumus Z-Score pada persamaan 2, karena data mempunyai satuan yang berbeda. Berikut adalah data yang telah distandarisasi.

Tabel 3. Data tabel 2 yang telah distandarisasi

Obyek (konsumen)	Usia (tahun)	Income (Rp)	Koran (Jam)	Tv (Jam)
A	−0,5636		1,1161	1,1555
B	−0,4025	1,5758	1,4733	0,5563
C		−0,8788	−0,6697	
D	1,8517	1,5758	0,759	−0,3424
E	1,0466	0,2121		−1,5407
F	0,2415	−0,061	−0,3125	−0,6419
G	−0,5636	−1,1515	−0,3125	
H		−0,3333		1,1555
I	−0,4025	−0,0606	−0,6697	
J	−1,2076	−0,8788	−1,384	−0,3424

Adapun langkah – langkah imputasi *missing data* dengan Algoritma *K-Means* :

1. Kelompokan seluruh obyek (termasuk obyek yang mengandung *missing data*) menggunakan Algoritma *K-Means*
   a. Tentukan jumlah *cluster* (k) dan jumlah iterasi maksimum (maxter). Misalkan k = 2 dan maxter = 100
   b. Tentukan *centroid* awal (C_0) pada obyek data komplit secara random misal :
      $v_1 = (-0{,}4025;\ 1{,}5758;\ 1{,}4733; 0{,}5563)$
      $v_2 = (-1{,}2076;\ -0{,}8788;\ -1{,}384;\ -0{,}3424)$
   c. Hitung jarak setiap obyek ke masing-masing *centroid* menggunakan rumus pada persamaan 3 :

$d(x_A, v_1) = 0{,}715865$ $\quad$ $d(x_A, v_2) = 2{,}98479$

$d(x_B, v_1) = 0$ $\quad$ $d(x_B, v_2) = 3{,}95533$

$d(x_C, v_1) = 3{,}258381$ $\quad$ $d(x_C, v_2) = 0{,}71432$

$d(x_D, v_1) = 2{,}529708$ $\quad$ $d(x_D, v_2) = 4{,}46949$

$d(x_E, v_1) = 2{,}890837$ $d(x_E, v_2) = 2{,}77623$

$d(x_F, v_1) = 2{,}778032$ $d(x_F, v_2) = 2{,}00181$

$d(x_G, v_1) = 3{,}263895$ $d(x_G, v_2) = 1{,}27955$

$d(x_H, v_1) = 2{,}0009$ $d(x_H, v_2) = 1{,}59409$

$d(x_I, v_1) = 2{,}696283$ $d(x_I, v_2) = 1{,}35197$

$d(x_J, v_1) = 3{,}955331$ $d(x_J, v_2) = 0$

d. Berdasarkan jarak terdekat setiap obyek ke masing-masing *centroid* didapat anggota *cluster* 1 yaitu obyek A, B, D dan *cluster* 2 yaitu obyek C, E, F, G, H, I, J.

e. Iterasi 1 :

Tentukan *centroid* pada iterasi 1 (C_1) menggunakan rumus pada persamaan 4 :

$v_1 = (0{,}2952; 1{,}5758;\ 1{,}1161;\ 0{,}4565)$

$v_2 = (-0{,}1771; -0{,}4502;\ -0{,}6697;\ -0{,}3424)$

Karena *centroid* pada iterasi 1 tidak sama dengan *centroid* awal ($C_1 \neq C_0$) atau karena t < maxter, maka hitung kembali jarak setiap obyek ke masing-masing *centroid*

$d(x_A, v_1) = 1{,}10727$ $d(x_A, v_2) = 2{,}36262$

$d(x_B, v_1) = 0{,}79017$ $d(x_B, v_2) = 3{,}09117$

$d(x_C, v_1) = 3{,}03543$ $d(x_C, v_2) = 0{,}42857$

$d(x_D, v_1) = 1{,}7856$ $d(x_D, v_2) = 3{,}20337$

$d(x_E, v_1) = 2{,}53233$ $d(x_E, v_2) = 1{,}83632$

$d(x_F, v_1) = 2{,}43477$ $d(x_F, v_2) = 0{,}73781$

$d(x_G, v_1) = 3{,}0788$ $d(x_G, v_2) = 0{,}87676$

$d(x_H, v_1) = 2{,}03304$ $d(x_H, v_2) = 1{,}50242$

$d(x_I, v_1) = 2{,}52063$ $d(x_I, v_2) = 0{,}45012$

$d(x_J, v_1) = 3{,}89512$ $d(x_J, v_2) = 1{,}32508$

Berdasarkan jarak terdekat setiap obyek ke masing-masing *centroid* didapat anggota *cluster* 1 yaitu obyek A, B, D dan *cluster* 2 yaitu obyek C, E, F, G, H, I, J

Iterasi 2 :

Tentukan *centroid* pada iterasi 2 (C_2) menggunakan rumus pada persamaan 4

$v_1 = (0{,}2952; 1{,}5758;\ 1{,}1161;\ 0{,}4565)$

$v_2 = (-0{,}1771; -0{,}4502;\ -0{,}6697;\ -0{,}3424)$

Karena *centroid* pada iterasi 2 sama dengan *centroid* pada iterasi 1 ($C_2 = C_1$), maka proses perhitungan dihentikan sehingga didapat anggota *cluster* 1 yaitu obyek A, B, D dan *cluster* 2 yaitu obyek C, E, F, G, H, I, J dengan *centroid*

$v_1 = (0{,}2952; 1{,}5758;\ 1{,}1161;\ 0{,}4565)$

$v_2 = (-0{,}1771; -0{,}4502;\ -0{,}6697;\ -0{,}3424)$

2. Isi *missing data* dengan *centroid* yang sesuai dengan letak *missing data* berada. Hasil imputasi dapat dilihat pada Tabel 4.

Tabel 4. Data hasil imputasi Algoritma *K-Means*

Obyek (konsumen)	Data Hasil Imputasi							
	Masih dalam standarisasi				Sudah dikembalikan kesatuan awal			
	Usia (thn)	Income (Rp)	Koran (Jam)	Tv (Jam)	Usia (thn)	Income (Rp)	Koran (Jam)	Tv (Jam)
A	−0,5636	1,5758	1,1161	1,1555	25	750000	10	20
B	−0,4025	1,5758	1,4733	0,5563	26	750000	11	18
C	−0,1771	−0,8788	−0,6697	−0,3424	27	300000	5	15
D	1,8517	1,5758	0,759	−0,3424	40	750000	9	15
E	1,0466	0,2121	−0,6697	−1,5407	35	500000	5	11
F	0,2415	−0,0606	−0,3125	−0,6419	30	450000	6	14
G	−0,5636	−1,1515	−0,3125	−0,3424	25	250000	6	15
H	−0,1771	−0,3333	−0,6697	1,1555	27	400000	5	20
I	−0,4025	−0,0606	−0,6697	−0,3424	26	450000	5	15
J	−1,2076	−0,8788	−1,384	−0,3424	21	300000	3	15

PENGUJIAN IMPUTASI *MISSING DATA*

Data yang digunakan dalam penelitian ini adalah data iris setosa yang diambil dari UCI *Machine Learning Reporsitory* [10]. Data tersebut merupakan data lengkap yang tidak terdapat *missing data*. Data iris merupakan kumpulan data multivariat yang diperkenalkan oleh Ronald Fisher. Data iris juga disebut dengan data iris Anderson, karena Anderson yang mengumpulkan data iris untuk mengukur variasi morfologi bunga iris dari tiga kelas yaitu iris setosa, iris virgina dan iris Versicolour. Masing-masing kelas terdiri dari 50 sampel. Empat fitur yang diukur dari masing-masing sampel adalah panjang dan lebar dari kelopak dan mahkota bunga. Pengukuran panjang dan lebar dalam cm. Dalam penelitian ini hanya digunakan satu kelas yaitu kelas iris setosa.

Skenario yang dilakukan pada penelitian ini adalah dengan menghilangkan beberapa nilai pada data iris setosa secara acak, dengan persentase *missing data* 10%, 20% dan 30%. Karena data iris tidak memiliki satuan yang berbeda maka data tidak perlu distandarisasi. Setelah dilakukan penghilangan nilai kemudian dilakukan imputasi *missing data* menggunakan metode *Mean* dan metode Algoritma *K-Means*. Setelah diperoleh hasil imputasi dari kedua metode, dilakukan evaluasi hasil imputasi dengan menghitung MSE. Berikut adalah nilai MSE dan nilai rata-rata MSE dari hasil imputasi kedua metode.

Tabel 5. Nilai MSE dan nilai rata-rata MSE dari hasil imputasi metode *Mean* dan metode Algoritma *K-Means*

Missing data	Metode		MSE					Rata-rata MSE
			Replikasi 1	Replikasi 2	Replikasi 3	Replikasi 4	Replikasi 5	
10%	***K-Means***	**2 *Cluster***	0,0861	0,0759	0,0322	0,0422	0,1051	0,0683
		3 *Cluster*	**0,0755**	**0,0724**	**0,0312**	**0,0251**	0,0709	**0,055**
		4 *Cluster*	0,1105	0,1037	0,0628	0,0261	**0,0686**	0,0743
	Mean		0,152	0,1363	0,0725	0,032	0,1515	0,1089
20%	***K-Means***	**2 *Cluster***	**0,0573**	**0,0661**	0,0527	**0,0605**	0,0885	0,065
		3 *Cluster*	0,0617	0,0677	**0,0498**	0,0639	**0,07**	**0,0626**
		4 *Cluster*	0,0723	0,0664	0,053	0,0659	0,0873	0,069
	Mean		0,1056	0,1121	0,0819	0,1025	0,1183	0,1041
30%	***K-Means***	**2 *Cluster***	**0,0537**	**0,05067**	0,0587	0,0546	0,0629	**0,0561**
		3 *Cluster*	0,0884	0,0609	**0,0493**	**0,0402**	0,0856	0,0649
		4 *Cluster*	0,0756	0,0811	0,0572	0,0514	**0,0532**	0,0637
	Mean		0,0896	0,0976	0,0874	0,0702	0,1009	0,0891

Berdasarkan Tabel 5 untuk persentase *missing data* sebesar 10%, hasil imputasi *missing data* menggunakan metode Algoritma *K-Means* dengan 2 *cluster*, 3 *cluster* dan 4 *cluster*, selalu memberikan nilai MSE yang lebih kecil dibanding metode *Mean* pada replikasi ke- 1, 2, 3 dan 5. Namun tidak pada replikasi ke-4, hasil imputasi menggunakan metode Algoritma *K-Means* dengan 2 *cluster* memberikan nilai MSE yang lebih besar dibanding metode *Mean* $(0{,}0422 > 0{,}032)$. Hal ini terjadi karena penghilangan nilai yang dilakukan secara acak, sehingga terkadang nilai-nilai yang hilang adalah nilai yang mendekati *mean* atau *mean* itu sendiri.

Untuk persentase *missing data* sebesar 20% dan 30%, hasil imputasi *missing data* menggunakan metode Algoritma *K-Means* dengan 2 *cluster*, 3 *cluster* dan 4 *cluster*, selalu memberikan nilai MSE yang lebih kecil dibanding metode *Mean* pada setiap replikasi. Berdasarkan nilai rata-rata MSE, hasil imputasi menggunakan metode Algoritma *K-Means* selalu memberikan nilai MSE yang lebih kecil dibanding metode *Mean* pada tiap persentase *missing data,* artinya secara rata-rata kesalahan hasil imputasi menggunakan metode Algoritma *K-Means* lebih kecil dibanding metode *Mean* sehingga secara rata-rata imputasi *missing data* menggunakan metode Algoritma *K-Means* menunjukkan hasil yang lebih baik dibanding metode *Mean.*

PENUTUP

Dari pengujian imputasi yang telah dilakukan, dapat disimpulkan bahwa secara rata-rata, imputasi *missing data* menggunakan metode Algoritma *K-Means* menunjukkan hasil yang lebih baik dibanding metode *Mean*. Untuk 10% *missing data*, hasil imputasi menggunakan metode Algoritma *K-Means* dengan 2 *cluster*, 3 *cluster* dan 4 *cluster* memberikan nilai MSE berturut-turut 0,0683; 0,055 dan 0,0743. Untuk 20% *missing data*, nilai MSE berturut-turut 0,065; 0,0626 dan 0,069. Untuk 30% *missing data*, nilai MSE berturut-turut 0,0561; 0,0649 dan 0,0637. Sedangkan hasil imputasi menggunakan metode *Mean* untuk 10%, 20% dan 30% memberikan nilai MSE berturut-turut 0,1089; 0,1041 dan 0,0891. Hasil imputasi menggunakan metode Algoritma *K-Means* memberikan nilai MSE yang lebih kecil dibanding metode *Mean*, karena nilai yang diimputasi oleh Algoritma *K-Means* merupakan nilai yang memiliki kemiripan dengan nilai-nilai yang berada dalam suatu *cluster*.

DAFTAR PUSTAKA

[1]. Santoso S. *Aplikasi SPSS Pada Statistik Multivariat*. Jakarta : PT Elek Media Komputindo; 2012.

[2]. Izzah A, Hayatin N. Imputasi Missing Data Menggunakan Algoritma Pengelompokan Data K-Harmonic Means. *Seminar Nasional Matematika dan Aplikasinya* (SNMA). 2013 sep 21.

[3]. Davey A, Savla J. *Statistical Power Analysis with Missing Data* [monograph online]. New York : Taylor and Francis Group; 2010 [cited 2014 Des 16]. Available from: Bookfi.org.

[4]. Li D, Deogun J, Spaulding W, Shuart B. Toward Missing Data Imputation: Study of Fuzzy K-Means *Cluster*ing Method. *Proceedings 4 th International Conference*. 2004 jun 1.

[5]. Enders CK. *Applied Missing Data Analysis* [monograph online]. New York : The Guilford Press; 2010 [cited 2014 Des 20]. Available from: Bookfi.org.

[6]. Acuna E, Rodrigues C. The Treatment of Missing Values and its Effect is the Classifier Accuracy. *Proceedings of the Meeting of the International Federation of Classification Sociaties (IFCS)*. 2004 jul 15.

[7]. Graham JW. *Missing Data Analysis and Design* [online]. USA: Springer; 2012 [cited 2014 Des 16]. Available from: Bookfi.org

[8]. Ediyanto, Mara MN, Satyahadewi N. Pengklasifikasian Karakteristik dengan Metode K-Means *Cluster* Analysis. *Bimaster*. 2013; 2(2):133-136.

[9]. Mehala B, Vivekanandan K, Thangaiah PRJ. An Analysis on K-Means Algorithm as an Imputation Method to Deal with Missing Values. *Asian Journal of Information Technology*. 2008; 7(9):434-441.

[10]. Asuncion A, Newman D. UCI Machine Learning Reporsitory [Internet]. 1988 [cited 2014 Des 20]. Available from: http://archive.ics.uci.edu/ml.

MUKARROMAH : FMIPA UNTAN, Pontianak, mukarromah123@yahoo.com
SHANTIKA MARTHA : FMIPA UNTAN, Pontianak, shantika.martha@gmail.com
ILHAMSYAH : FMIPA UNTAN, Pontianak, ilhamsm99@gmail.com